Bayesian Analysis of COVID-19 cases in California

Gulzina Kuttubekova[*]

Abstract

Since mid-January of 2020, we have seen the first cases of coronavirus disease (COVID-19) in the US. Serious respiratory disease caused by the novel virus had taken the lives of approximately 300,000 people in the world and infected more than 4.3 million people worldwide. `https://www.worldometers.info/coronavirus/` provides various COVID-19 related statistics about the US as well.

We examine the dataset of COVID-19 cases in California, along with the death rate among those who were found infected by this novel virus. We also look for a possible correlation between the number of infected people and the population in each county. We employ two hierarchical models to examine the infection rate and assess models by posterior predictive checks.

A modified hierarchical model on the joint distribution of the number of infections and the number of deaths showed a better fit than the hierarchical model which doesn't take into account the mortality rate and its distribution. In general, we saw systematic differences in the number of deaths and infections across counties. The assumption that the mean number of the number of infections is 20% of the population had a high impact on the posterior distribution in the first model.

Key Words: Bayesian analysis, COVID-19, hierarchical models, MCMC

1. Introduction

The novel virus is still spreading across the world despite the precautions taken by governments of different countries. The numbers connected to this pandemic are changing very rapidly, not every day, but every hour. There are only 13 countries left where there are no officially confirmed cases of COVID-19, according to `https://www.aljazeera.com/news/`. However, it's believed that it's only a matter of time while the novel virus reaches those countries. Since the first massive cases of COVID-19 and related death numbers started increasing in the Greater Seattle area in the US in mid-March, the epicenter of outbreak shifted quickly from West to East coast. At the current time, the Greater New York area became the epicenter of coronavirus battle. Although the state of California has as twice as the population as New York state, according to Worldometers number of reported infected people in the New York area is seven times greater than in California. Obviously there are many external factors that can explain this discrepancy.

We want to know if there exists any discrepancy in the distribution of COVID-19 cases among the counties in California. If yes, which counties are more influential in explaining the number of death per reported number of infected people. We are also interested in the variation in the number of infected people by county. It's expected that as larger the population of the county as greater the number of infected people. However, it might be a false expectation, since there are other factors influencing the infection rate as population density, social culture, access to healthcare services and etc. We also integrate the new assumption that infection depends on mortality to answer the proposed research questions.

We employ two hierarchical Poisson models constructed in different ways. We tackle the problem stated above using a fully Bayesian approach. Results

[*]UCSC Baskin School of Engineering, Department of Statistical Science

from those models are reported and compared. We also use results from the previous study, which are available at `https://github.com/kgulzina/bayesian-analysis-of-COVID19-ir` for additional model checks.

1.1 Data

We analyze the dataset which contains information about incidence and mortality due to coronavirus in California. The number of reported cases and death is given per each county as of 04/13/2020. There are 58 observations, as there are 58 counties in the state in total. There are 4 variables: $\{$County, Total.cases, Deaths, Population$\}$ s.t. population, the number of infected people, and number of people who died from COVID-19 are given. As we have count variables except for the county name, we treat the number of deaths y_i out of n_i infected people per i^{th} county with population c_i as discrete random variables. We also treat y_i and n_i as response variables.

1.2 EDA

The top three counties with the largest number of infected people are Los Angeles, San Diego, and Riverside. Although San Francisco county has the largest population density $\sim 18,553$, among all 58 counties in the state, it has 11 times less number of infected people compared to Los Angeles county with $\sim 2,488$ population density. There are five counties that have zero reported cases of COVID-19. All five counties have population $\leq 32,000$. Interestingly four of those counties are in Northern California, and Mariposa county is relatively close to the highly populated county like Santa Clara. Up to date, $\sim 0.06\%$ of the population of California were infected with the virus. We would like to learn the trend in the number of deaths as this percentage increases. Moreover, throughout the study, we assume the mean number of coronavirus cases is equal to 20% of the population.

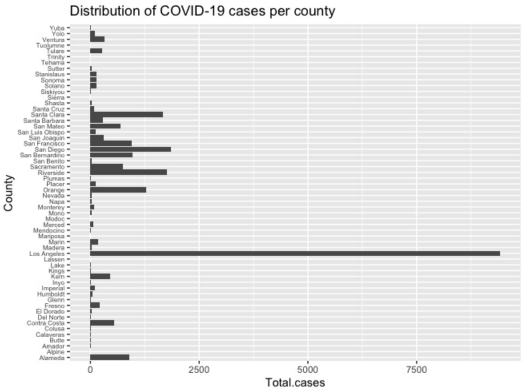


Figure 1: Distribution of the total number of confirmed cases of COVID-19 across counties of California.

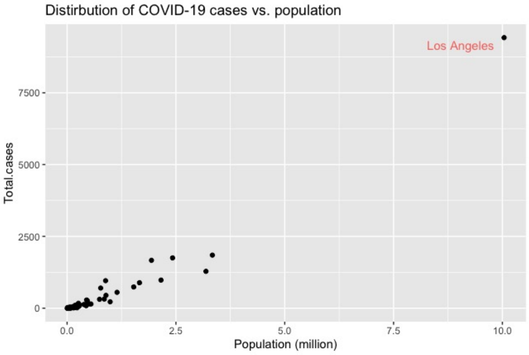
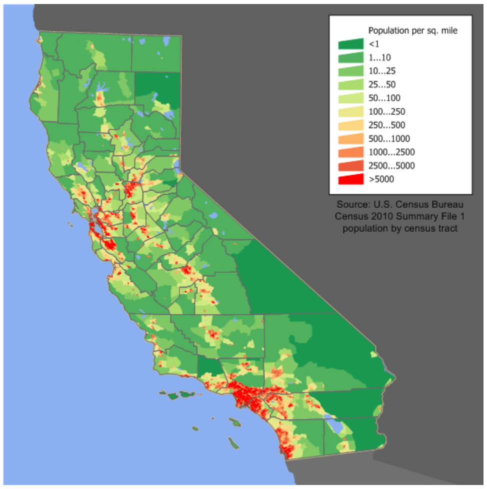


Figure 2: Scatter plot number of confirmed cases of COVID-19 vs population for each county in California.

According to the plot in Figure 1, we see that number of confirmed cases is not evenly distributed across counties. Los Angeles has the most noticeable spike in cases. The top five counties with the total number of infected people $\geq$ 1,000 are Los Angeles, San Diego, Riverside, Santa Clara, and Orange counties. Apparently these high numbers are associated with the population in each county. Figure 2 shows that as population number increases the number of infected people increases as well. Scatter plot shows there is a potential outlier: Los Angeles county, which has the largest population and the highest number of infected people as well. As we stated earlier, there might be some other external factors affecting the non-homogenous distribution of cases across counties. For instance, infection numbers can be related to population density as we see in Figure 3, the top 5 in-



Source: U.S. Census Bureau
Census 2010 Summary File 1
population by census tract

Figure 3: Population density of California (Hines 2019).

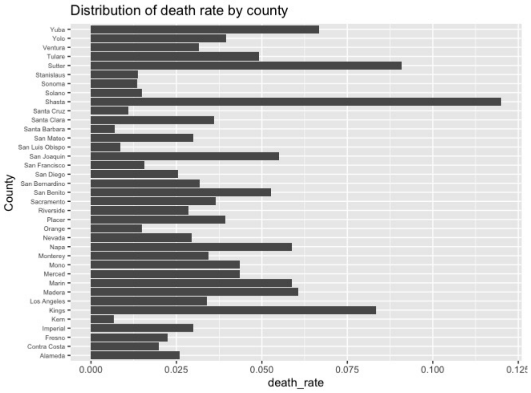


Figure 4: Distribution of death rate due to COVID-19 across counties of California.

fected counties have population density of $\geq$ 1,000. Note that in the previous study we treated the number of confirmed cases n_i as a known quantity, and now we treat it as a random variable of interest.

Figure 4 shows that the death rate, i.e. the proportion of deaths per confirmed coronavirus cases are far from being the same. For instance, Shasta, Sutter, and Kings counties have a greater mortality rate than in Los Angeles county. Even if we see this discrepancy, we can argue and treat the death rate to be the same across counties. The answer is obvious: since we don't have enough observations of confirmed cases in small populated counties like Sutter ($\leq 100,000$) empirical estimate of proportion may give misleading information about the true death rate. For instance, assume there is a county that has only 10 confirmed cases of COVID-19, by chance 2 people have serious underlying illnesses and those two patients die. Then the empirical estimate of proportion is 0.2! Which is greater than the sample death rate in any other county.

Our preliminary findings suggest that the distribution of total cases of coronavirus differ by counties. Although we assume that the death rate is the same across counties, we test it by employing different models. In the next section, we describe models in detail.

2. Methods

As stated before, number of death y_i out of n_i infected people per county i are independent observations from Binomial distribution conditioning on $\theta_i = P(death)$. Also, population of each county c_i is considered as an explanatory variable for modeling distribution of number of infections n_i.

2.1 Hierarchical Poisson model with Gamma priors

First, we consider a model:

$$n_i \sim Poi(\lambda_i c_i/10^3)$$

$$\lambda_i \sim Ga(\alpha, \beta)$$

$$p(\alpha, \beta) = Ga(\alpha|a_\alpha, b_\alpha)Ga(\beta|a_\beta, b_\beta)$$

where prior and hyperprior distributions have $Ga(shape, scale)$ structure and all parameters are greater than zero.

We set the values of hyperparameters of the hyperprior distribution $p(\alpha, \beta)$ according to the assumption that, for a given county, the mean number of cases is equal to 20% of the population. To include this assumption we do the following procedure:

1. For county i,

$$E[n_i|\lambda_i] = \lambda_i c_i/10^3 = 0.2c_i$$

$$\lambda_i = 200$$

2. Assuming that distribution of λ should be concentrated around 200, we set
$$E[\lambda_i|\alpha,\beta] = \alpha\beta = 200$$
3. Then we let
$$E[\alpha|a_\alpha, b_\alpha] = a_\alpha b_\alpha = 20$$
$$E[\beta|a_\beta, b_\beta] = a_\beta b_\beta = 10$$
4. Consequently, we set $a_\alpha = 10, b_\alpha = 2, a_\beta = 20, b_\beta = 0.5$.

We write the posterior distribution as

$$p(\lambda,\alpha,\beta|\mathbf{n}) = p(\lambda|\alpha,\beta,\mathbf{n})p(\alpha,\beta|\mathbf{n})$$

We can also write $p(\lambda,\alpha,\beta|\mathbf{n})$:

$$\propto p(\alpha,\beta)p(\lambda|\alpha,\beta)p(\mathbf{n}|\lambda)$$
$$\propto p(\alpha,\beta)\prod_{i=1}^{58} Ga(\lambda_i|\alpha,\beta)Poi(n_i|\frac{\lambda_i c_i}{10^3})$$

Using this joint posterior we get the full conditional distribution of $\lambda : p(\lambda|-)$

$$p(\lambda|\alpha,\beta,\mathbf{n}) = \prod_{i=1}^{58} p(\lambda_i|\alpha,\beta,n_i)$$

$$p(\lambda_i|\alpha,\beta,n_i) = Ga(\alpha+n_i, (\frac{1}{\beta}+\frac{c_i}{10^3})^{-1})$$

Now we have to find $p(\alpha,\beta|\mathbf{n})$:

$$p(\alpha,\beta|\mathbf{n}) \propto p(\alpha,\beta)p(\mathbf{n}|\alpha,\beta)$$

where $p(\mathbf{n}|\alpha,\beta) = \prod_{i=1}^{58} p(n_i|\alpha,\beta)$ and

$$\begin{aligned} p(n_i|\alpha,\beta) &= \int p(n_i|\lambda_i)p(\lambda_i|\alpha,\beta)d\lambda_i \\ &= \int Poi(n_i|\frac{\lambda_i c_i}{10^3})Ga(\lambda_i|\alpha,\beta)d\lambda_i \\ &= \frac{\tilde{c}_i^{n_i}\Gamma(\alpha+n_i)}{n_i!\Gamma(\alpha)\beta^\alpha}(\frac{1}{\frac{1}{\beta}+\tilde{c}_i})^{(\alpha+n_i)} \end{aligned}$$

where $\tilde{c_i} = c_i/10^3$.

We use this factorization to sample from the joint posterior distribution. First, we use `sample()` function in R and Griddy sampling approach to sample from $p(\alpha,\beta|\mathbf{n})$ and using those samples, we directly draw λ_i from $Ga(\lambda_i|-)$ for each $i = 1,...,58$. Further, we get point and interval estimates of posterior mean and posterior variance using that sample for each parameter.

We also obtain samples from the predictive posterior distribution of $\mathbf{n}$ in order to compare the predictive distribution of the number of deaths per 1000 habitants for the different counties, in combination with samples of $\theta = Prob(death)$ from the previous study.

We can obtain posterior predictive simulations of $\tilde{n}$ from 58 counties:

- draw (α,β) from $p(\alpha,\beta|\mathbf{n})$
- draw 58 parameters from $p(\tilde{\lambda}_i|\mu,\tau)$
- for each c_i in $i = 1,...,58$, draw $\tilde{n}_i$ from $\tilde{n}_i \sim p(\tilde{n}_i|\lambda_i c_i/10^3)$

2.1.1 *Predictive distribution of the number of deaths per 1000 habitants for different counties*

Consider the Beta-Binomial hierarchical model from the previous study:

$$y_i \sim Bin(n_i,\theta_i), \theta_i \sim Be(\mu\tau,(1-\mu)\tau)$$
$$p(\mu,\tau) = (\mu(1-\mu)(1+\tau)^2)^{-1}$$

We can write the posterior distribution of θ,μ,τ as follows:

$$p(\theta,\mu,\tau|\mathbf{y}) = p(\theta|\mu,\tau,\mathbf{y})p(\mu,\tau|\mathbf{y})$$

We draw samples from the posterior distribution using the factorization above. I.e. first we draw samples of μ,τ from $p(\mu,\tau|\mathbf{y})$, then using these samples we draw samples of θ from $p(\theta|\mu,\tau,\mathbf{y})$.

Full details of derivation of posterior distribution for parameters θ,μ,τ are given in Kuttubekova's (2020) report. We sample μ,τ using *rejection sampling* and θ_i using *direct sampling*.

We estimate mean posterior death rate for each county $E[\theta_i|y_i]$ using samples from joint posterior, and use those estimates to obtain samples from posterior predictive distribution of the number of deaths. For each c_i and $\tilde{n}_i = (\tilde{n}_{1,i}, \tilde{n}_{2,i}, ..., \tilde{n}_{n,i})$ s.t. $i \in 1, ..., 58$ and n is the sample size of $\tilde{n}_i$, we have $\tilde{y}_i = \tilde{n}_i \hat{E}[\theta_i|y_i]$.

2.2 Hierarchical Poisson model with Gamma priors (Modified)

Now we consider a modification of the previous model where infection rate depends on the mortality rate:

$$f(y_i, n_i|\theta_i, \lambda) = f(y_i|n_i, \theta_i, \lambda)f(n_i|\theta_i, \lambda)$$
$$= Bin(y_i|n_i, \theta_i)Poi(n_i|\theta_i, \frac{\lambda c_i}{10^3})$$

for $\theta_i \in (0, 1)$ and $\lambda > 0$. We also assume that $\theta_i \sim Be(\theta_i|\mu\tau, (1-\mu)\tau)$ and we set priors:

$$p(\mu, \tau) = (\mu(1-\mu)(1+\tau)^2)^{-1}$$

where $\mu \in (0, 1)$, $\tau > 0$ and

$$\lambda \sim Ga(\lambda|a, b)$$

where a and b are fixed. We set values of $a = 20$ and $b = 10$ as in the previous case, according to the same assumption that the mean number of cases is equal to 20% of the population. It's might not the best values for hyperparameters, however, we will check if the model is sensitive to our choice of a and b later in the report.

We can write joint posterior distribution of θ, λ, μ and τ as follows:

$$\begin{aligned} p(\theta, \lambda, \mu, \tau|\mathbf{y}, \mathbf{n}) &\propto p(\theta, \lambda, \mu, \tau) \\ &\times p(\mathbf{y}, \mathbf{n}|\theta, \lambda, \mu, \tau) \\ &= p(\lambda, \mu, \tau)p(\theta|\lambda, \mu, \tau) \\ &\times p(\mathbf{n}|\theta, \lambda, \mu, \tau) \\ &\times p(\mathbf{y}|\mathbf{n}, \theta, \lambda, \mu, \tau) \\ &= p(\mu, \tau)p(\lambda)p(\theta|\mu, \tau) \\ &\times p(\mathbf{n}|\theta, \lambda)p(\mathbf{y}|\mathbf{n}, \theta, \lambda) \end{aligned}$$

Using this explicitly written joint posterior distribution, we can find full conditional distributions of each parameter up to a constant. Note that we are interested in posterior samples of parameters θ and λ. However, we use μ, τ as latent variables, so it's easier to sample from the joint posterior.

$$p(\lambda|-) = Ga(\lambda|a + \sum_{i=1}^{58} n_i, (\frac{1}{b} + \sum_{i=1}^{58} \frac{\theta_i c_i}{10^3})^{-1})$$

And for $i \in \{1, ..., 58\}$:

$$\begin{aligned} p(\theta_i|-) &\propto exp(-\theta_i \frac{c_i}{10^3})\theta_i^{\mu\tau+n_i+y_i-1} \\ &\times (1-\theta_i)^{(1-\mu)\tau+n_i-y_i-1} \end{aligned}$$

Finally,

$$p(\mu, \tau|-) \propto p(\mu, \tau) \prod_{i=1}^{58} Be(\theta_i|\mu, \tau)$$

Note that the last full conditional is also proportional to

$$p(\mu, \tau) \prod_{i=1}^{58} Be(\theta_i|\mu, \tau)Bin(y_i|n_i, \theta_i)$$

which is proportional to the $p(\mu, \tau|\mathbf{y})$ from Beta-Binomial model if we integrate out θ_i's. For simplicity, we will use the latter kernel to obtain samples from posterior joint distribution.

Since we were able to find conditional distributions of each parameter, we employ *Gibbs sampling* to sample from posterior. Full conditionals are not given in the closed form, except the distribution for λ. So we employ *Metropolis-Hastings* algorithm within the Gibbs algorithm to sample θ_i's. We set $N(\theta_i^s, 0.01)$ as a proposed distribution for each $p(\theta_i|-)$ inside Metropolis Hasting algorithm. Also, we use *rejection sampling* within Gibbs algorithm to sample μ, τ. Note that the whole process of setting proposal distribution for rejection sampling is already given in Kuttubekova (2020).

After we obtain samples, we report point and interval estimates of the posterior mean and posterior variance for each parameter. We also compare the distributions of mortality rate i.e θ_i for each county.

We also obtain samples of $\tilde{n}_i, \tilde{y}_i$ from posterior predictive distribution, and compare its distribution across counties. Note that we use the same samples to do posterior predictive checks in order to assess model fit.

2.3 Model assessment

Posterior predictive checks can be done in many different ways, but mainly all those methods try to evaluate if the phenomenon seen in observed data is also captured and reproduced by the model.

From EDA there are some unusual findings in observed data. For instance, San Francisco county with 884,363 population has almost the same number of infected people as those with a substantially higher population. Orange and San Diego counties have almost the same population densities, but San Diego has 1.5 times more infected people than Orange county. We want to use this phenomenon in order to assess model fit. We define test quantity

$$T(n_i, \lambda_i) = \frac{n_i}{\sum n_i}$$

= proportion of number of infected for county i over the total number of infected people in California.

We also employ classical DIC and Gelfand & Ghosh (1998) criteria to compare model fits. Note that the model is favorable if it has a smaller DIC and Gelfand & Ghosh criterion. BIC and AIC cannot be used here, since they don't account for hierarchical models and they directly evaluate log-likelihood at MLE estimates of parameters, rather than at MAP estimates.

We are also interested in the effect of the hyperparameter values a and b on the overall model fit. To do sensitivity analysis, we fit the model over a grid of values for hyperparameters. Assessing criteria would be the posterior predictive distribution of the number of infected people in Orange county. (Orange county is chosen at random from the list of counties with some phenomenon).

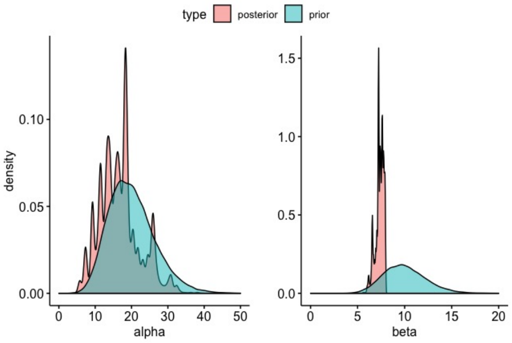


Figure 5: Prior and posterior distribution of : α and β parameters

Table 1: Posterior mean and variance estimated by Model 1.

Parameter	2.5%	50%	97.5%
α	7.296	16.252	29.445
β	6.272	7.410	7.950

3. Analysis

3.1 Hierarchical Poisson model with Gamma priors

We obtain 50,000 sample draws from the joint posterior distribution of λ, α, and β. As we see in Figure 5, the posterior distribution of α is updated but still on the same range as its prior. However, the posterior distribution of β is more concentrated around the left quarter tail of its prior distribution. Note that the model was given two hyperprior distributions for α and β s.t. they account for the assumption that on average 20% of the population in each county contracts coronavirus. We see that this assumption and the Poisson likelihood added their contribution in generating joint posterior and updating prior.

Point and interval estimates of α and β are given in Table 1. The same shrinkage of posterior distribution of β can be seen from 95% credible interval.

Now we analyze a sample from posterior distributions of each λ_i for 58 counties. Figure 6 shows the posterior distributions of each λ_i for $i = 1, ..., 58$ depicted by box plots. Distribution for Los Angeles is one of

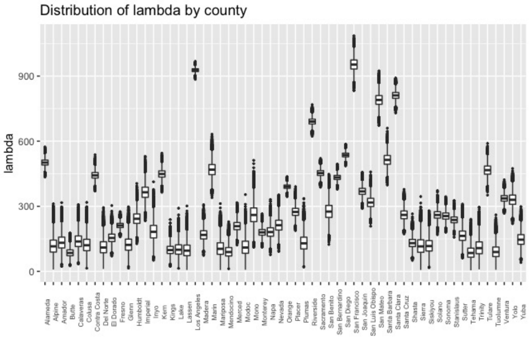


Figure 6: Posterior distribution of λ_i's depicted by box plots.

the highest and it's also shrunk, comparative to the distribution for San Francisco county, which is of the same height but has more uncertainty. The sampling distribution of n_i is $Poi(\lambda_i c_i/10^3)$, so that λ_i serves as coefficient parameter for each a scaled covariate $c_i/10^3$. For instance, Los Angeles county has population $c_i = 10,039,108$ and hence estimated $E[n_i|\lambda_i]$ is 9,546,425. It means assumption, that 20% of the population will be infected, combined with observed data estimates infection of 95% of the Los Angeles population. We also estimate the same rate $\sim 92\%$ for San Francisco county. These estimates are suspiciously high, it might be a lack of fit of data to a model, or we'll to see such high numbers of infection in those counties in the future.

Overall, there is a noticeable difference in posterior distributions of λ_i's across counties. Top five counties with the highest quartile 1 in the distribution of λ_i's are Los Angeles, San Francisco, San Mateo, Santa Clara, and Riverside. Note that Orange and San Diego counties were among the top five counties with the highest number of infected people per county according to observed data.

We draw 1,000 samples from the posterior predictive distribution of n_i for each county. We included the samples of θ_i's from the previous study's hierarchical model to obtain samples of predictive distribution of the number of deaths. If the model fits the data well, it should be able to replicate data. We noticed in EDA that Los Angeles county has the largest proportion of infections across counties in California. We use this quantity as $T(\mathbf{n}, \lambda)$ and calculate the probabilities of observing more extreme values of it from replicated data for different counties. Counties in Figure 7 mostly have some discrepancies from the general trend which is confirmed by posterior probabilities in Table 4.

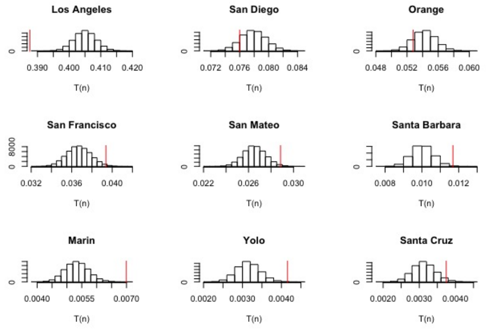


Figure 7: Distribution of test quantity on replicated data for different counties.

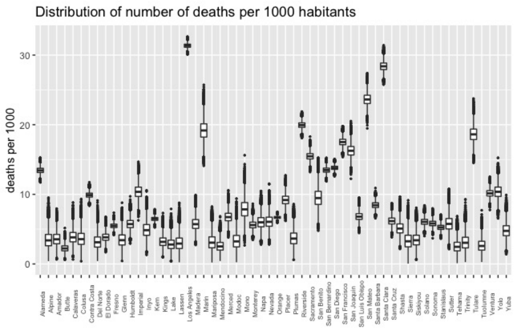


Figure 8: Posterior predictive distribution of the number of deaths per 1000 habitants for 58 counties.

3.1.1 Predictive distribution of the number of deaths per 1000 habitants for different counties

We obtain 50,000 samples of $\tilde{n}_i$ from $p(\tilde{n}_i|n)$ for each county. We also retrieved sample results from previous study: 250 samples of θ_i from $p(\theta, \mu, \tau|y)$. We combine those sampled to obtain 50,000 samples of the number of deaths per 1000 habitants for each county. Figure 8 shows that the distribution of the number of deaths across

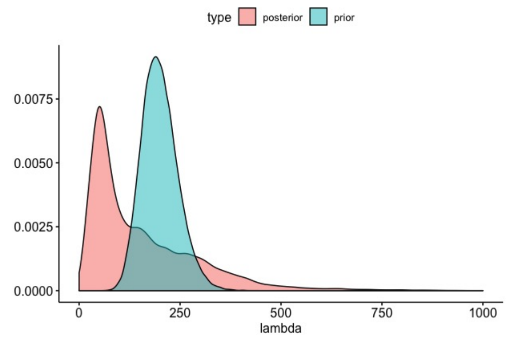


Figure 9: Posterior distribution of λ

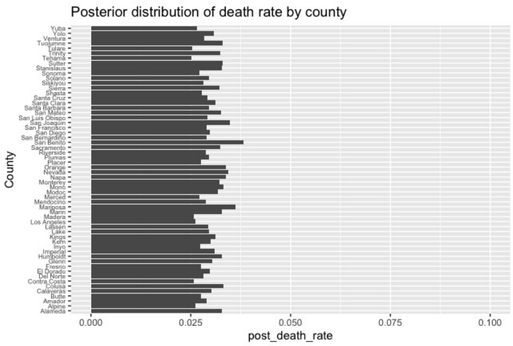


Figure 10: Posterior distribution of mortality rate

counties is not homogenous. There are 4 counties with the median number of deaths per 1000 habitants $\geq$ 20 and Los Angeles county with the median ≥ 30. I.e, more than 30 people per 1000 die in Los Angeles due to a novel virus.

Table 3 shows that more than the number of deaths changed tremendously, as it was impacted by the assumption we gave in prior distribution.

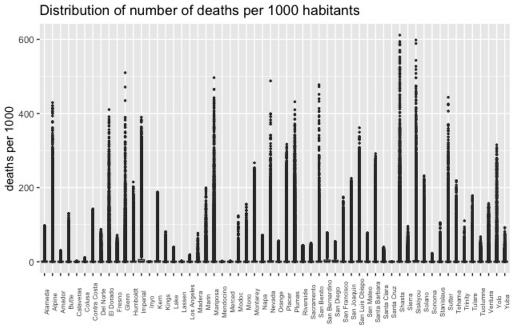


Figure 11: Posterior predictive distribution of number of deaths per 1000 habitants.

3.2 Hierarchical Poisson model with Gamma priors (Modified)

We obtain 50,000 samples from the joint posterior distribution of θ, λ, μ and τ. Although we are interested in samples of θ and λ, we used μ, τ as latent variables for an easier sampling procedure from the joint posterior distribution.

Figure 9 shows the prior and posterior distributions of λ in model 2. The posterior distribution is highly skewed to the right, while prior is more concentrated. The posterior mean of λ can be easily calculated since we know its full conditional distribution in a closed form. Hence, $E[\lambda|\mathbf{y}, \mathbf{n}] = 154.75$ and 95% credible interval is $[19.46, 492.97]$. Note that credible is quite wide, which is probably because of a heavy right tail.

According to the Figure 10 the distribution of mortality rate has been homogenized comparative to its distribution in Figure 4. We set non-informative $Beta(0,0)$ prior for μ, where $\theta_i \sim Be(\theta_i|\mu, \tau)$. So posterior of θ_i is dominated by the Beta likelihood with combination with data likelihood, which also involve parameter θ_i.

We obtain 50,000 samples of $\tilde{n}_i$ and $\tilde{y}_i$ for each county. Using those estimates, we obtain the posterior predictive distribution of the number of deaths. Figure 11 shows that the distribution of the number of deaths per 1000 habitants is far from being homogenous. We see that almost all distributions are highly skewed with the right heavy tails. We refer to Table 3 and see that surprisingly Los Angeles has fewer deaths per 1000 than other counties..

We have seen in EDA that the distribution of the number of infections is heterogenous, i.e each county has significantly different (small or high) numbers. We would like to know if this trend is seen in predictions made by model 2. Figure 12 shows that the distribution is uniform. However, some counties like Imperial, Solano, San Luis Obispo, which had a smaller number of infections tend to have slightly more than the average.

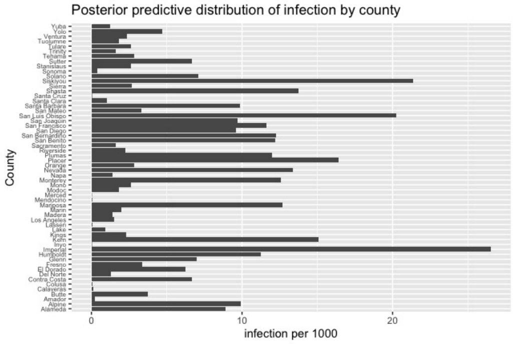


Figure 12: Posterior predictive distribution of number of infections per 1000 habitants.

Table 2: Posterior probability of observing extreme T(n) for Orange county. Calculated by model 2

a	b	T(n)
20	10	0.1342
20	5	0.1265
5	5	0.4226
0.5	0.5	0.4927

Table 3: Distributions of the number of deaths per 1000 habitants for each county. Estimated by model 1 & 2

County	$E_1[\tilde{y}\|y]$	$E_2[\tilde{y}\|y]$	y
Los Angeles	31.35	0.610	0.032
Santa Clara	28.39	0.402	0.031
San Mateo	23.64	1.309	0.027
Riverside	19.95	1.270	0.021
Marin	19.20	0.824	0.038

Table 4: Posterior probability of observing extreme T(n) calculated by model 1 & 2

County	Model1	Model 2
Los Angeles	0.0000	0.0883
San Diego	0.1368	0.3887
Orange	0.1648	0.1342
San Francisco	0.0130	0.2225
San Mateo	0.0107	0.1146
Santa Barbara	0.0028	0.2313
Marin	0.0001	0.0517
Yolo	0.0009	0.1373
Santa Cruz	0.0299	0.0009

As you see from the table our quantity of interest is affected by a change in the values of a and b. As their values decrease we see better fit (in terms fo Orange county only). So we have to be very careful in the choice of hyperparameter values. We can evaluate sensitivity w.r.t. other test quantities or aspects of the data.

3.3 Comparing two models

We fit five models in total to the COVID-19 data. The last three models are hierarchical, and the last two treat the total number of infected people n_i in each county as a random variable rather than fixed. The last model also takes into account potential overdispersion in the number of deaths.

We see in Table 3 the predictive distribution by model 1 was influenced by the assumption that the mean number of cases is equal to 20% of the population. Model 2 has also overestimated the mean number of deaths comparative to a summary from observed data. On average, it's expected that the number of deaths per 100 habitants is between 5 and 14 (for counties in the table) by model 2, while the number of deaths per 1000 habitants is between 20 and 31 by model 1.

We draw samples from posterior predictive distribution of the number of infected people in each county to see if models were able to replicate the observed data. Initially, we observed that Los Angeles county had the highest number of infected people across all 58 counties, and we want to know if the same trend is the case for both models. Posterior probabilities of observing as extreme values as a test quantity are given in Table 4 for both models. Note that we also test other counties, which have unusual behavior in the distribution of the number of infected

Table 5: Model fit assessment

Model	DIC	Gelfand-Ghost
Model 1	294.13	1.005e+14
Model 2	-9182.21	1.321e+10

people. Posterior probabilities calculated by model 2 are slightly better than in model 1. However, there is no major improvement except for counties like San Diego, San Francisco, and Santa Barbara.

We employ two different methods to compare the models, and probably find a better fit. We also used two criteria to choose between models. Gelfand & Ghosh criteria compare the predictive abilities of two models. Although the criteria values are very high for both models, the second modified model has substantively smaller criteria value. Also, we employed DIC, which is a generalization of AIC for hierarchical models. It's calculated by estimating the effective number of parameters in the model. We know that negative DIC values are legit and small values of DIC are preferred. Hence, by both criteria model 2 seems to be a better fit for the COVID-19 data than model 1.

There was a tremendous difference in the computation time of two models for obtaining samples from posterior. I.e model 2 required 3 times more time than model 1 for the same number of samples. Since we employed a combination of Metropolis-Hastings algorithm and rejection algorithms for model 2, it required more time to run 5 chains and obtain 50,000 samples from posterior. That time requirement triples since we obtain samples of $\tilde{n}_i$ and then samples of $\tilde{y}_i$ for each county.

Although the second model is more complex, it was easier to obtain full conditionals and construct a working sampling algorithm. We could use Gibbs sampler for the first model as well if we treated λ_i's as a latent variable. Our approach was to use the Griddy approach to sample from $p(\alpha, \beta | mathbfn)$ which created a lot of computational issues so we manually simplified the kernel of $p(\alpha, \beta | mathbfn)$ and wrote many additional functions to evaluate density only.

4. Discussion

We employed two different constructions of the Poisson-Binomial models. We treated the number of infected people as a random variable along with the number of deaths per county. Our preliminary results showed that the infection rate and death rate across counties are not homogenous. It was also confirmed throughout by fitting the two models above.

We assumed that 20% of people in California will contract the virus. Actually we used it as a piece of prior information in the first model and used it indirectly in the second model as well. We ignored the time feature of the spread of the virus, i.e we are not interested in estimating when will those 20% be infected. So it can be the next step to include time in our model.

As a next step, we can consider a model that accounts for the time and space varying nature of the spread of the virus. I.e we might want to employ Gaussian or Poisson processes that are suitable for time series as well as spatial data.

REFERENCES

Aljazeera, News https://www.aljazeera.com/news/

Hines B 2019, Population Density of California https://storymaps.arcgis.com/stories/252b65cbe331420389c61453c61ea221

Kuttubekova G 2020, Bayesian Analysis of COVID-19 cases in California https://github.com/kgulzina/bayesian-analysis-of-COVID19-in-CA

Worldometer: Coronavirus, https://www.worldometers.info/coronavirus/

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013), *Bayesian Data Analysis*(3^{rd} ed.). CRC Press. Taylor & Francis Group.